Three Variations of Genetic Algorithm for Searching Biomolecular Conformation Space: Comparison of GAP 1.0, 2.0, and 3.0

A. Y. JIN, F. Y. LEUNG, D. F. WEAVER

Department of Chemistry, Queen's University, Kingston Ontario K7L 3N6, Canada

Received 3 March 1998; accepted 22 February 1999

ABSTRACT: Three genetic algorithm programs, GAP 1.0, 2.0, and 3.0, were used in conjunction with the ECEPP/2 force field to search the conformation space of [Met]-enkephalin. Each program was proficient at quickly finding many diverse low-energy conformers. Conformer populations displayed a variety of secondary structure motifs including those likely to bind to the μ-opioid receptor. Limitations in the program's sampling behavior are discussed and method improvements are suggested. Although still in a developmental stage, the GAP programs represent a useful addition to conformational search techniques when no *a priori* structural information is available. © 1999 John Wiley & Sons, Inc. J Comput Chem 20: 1329–1342, 1999

Keywords: genetic algorithm; ECEPP/2; conformational search; [Met]-enkephalin; peptide

Introduction

Theoretically elucidating the three-dimensional structures of biological macromolecules is a fundamental problem in computational chemistry; a practical solution remains elusive because of the "multiple minima problem."[1,2] This problem is a manifestation of the enormous number of local minima on the potential energy hypersurface of a conformationally flexible molecule. Systematically searching such a hypersurface is rendered impractical because the elucidation of low-energy conformations, including that at the global minimum energy, must be achieved by nonanalytical, computationally expensive means. For biomolecules such as linear peptides, this problem is especially relevant because conformational diversity plays a pivotal role in interactions with multiple receptor proteins. Thus, the view of the multiple

Correspondence to: D. F. Weaver, e-mail: weaver@chem.queensu.ca

Contract/grant sponsor: Ontario Ministry of Health

CCC 0192-8651 / 99 / 131329-14

minima problem as an extremely difficult energy minimization task can be extended beyond identifying the global minimum and toward determining families of biologically relevant low-energy conformational domains. Such a conceptualization is important for many peptides such as the endogenous opioid [Met]-enkephalin pentapeptide. From three different x-ray diffraction studies it has been shown that this peptide adopts an extended conformation in the solid state.[3–5] In contrast, proton NMR studies reveal a propensity for a β-turn structure as well as the absence of a single well-defined solution-phase conformation.[6,7] Pharmacological studies reveal that differing conformations interact with different receptors. Clearly, it cannot be assumed that the global minimum-energy conformation of [Met]-enkephalin is sufficient to establish a meaningful structure-activity hypothesis under physiological conditions. It is imperative to devise methods that reveal diverse low-energy conformer populations reflecting a structural spectrum of biologically relevant states.

Although molecular dynamics and Monte Carlo simulations have traditionally been used to search macromolecular conformational space, genetic algorithms (GAS) are emerging as a useful approach. As a relatively new class of stochastic optimization methods, GAs[8–11] have yielded interesting results in numerous applications ranging from medical bioinformatics[12] to airframe design.[9] Within the past 5 years, GAs have also seen increasing use among diverse problems of chemical interest: instrument configuration; chemometric analysis; spectral analysis; the design of combinatorial libraries; and the refinement of NMR solution structures.[13–21] Topics that have received intensive scrutiny include protein folding,[22–32] drug-receptor docking,[33–36] and the structure prediction of both small molecules and large macromolecular assemblies.[37–45] Although the performance of GAs throughout chemistry is still in the early stages of evaluation, the relative novelty of this search paradigm has enticed widespread attention. The GA method is based on principles that have been gleaned from the study of adaptive learning in both natural and artificial settings. Utilizing processes associated with evolution and gene recombination as metaphors for data-manipulating operations, GA-driven conformational search methods present an interesting compromise between the stochastic exploration of conformer space and the exploitation of previously generated conformer information.

Although there have been many reports of successful GA applications in global optimization tasks, the *a priori* parameterization of GA methods remains a trial-and-error process. Moreover, for GA-based conformational searches of flexible molecules, practical limitations often require the introduction of ad hoc modifications. For example, the use of a finite population size and a finite number of generations introduces statistical error, which accumulates over the GA run. This has led to modifications motivated by expediency rather than mathematical rigor. In a previous study, we examined some of these issues by developing the GAP 1.0 program,[46] which utilizes the ECEPP/2 force field.[47–50] To adapt the genetic algorithm to the identification of low-energy conformation space, GAP 1.0 used a ''uniform crossover'' operator and a ''diversity'' operator. These were designed to deal with the ''preconvergence'' problem that can occur in small populations. Although GAP 1.0 presented a simple GA-driven optimization scheme, it proved adept at exploring low-energy conformation space. The current investigation extends GAP 1.0 to two additional versions. In addition to the uniform crossover operator used previously in GAP 1.0, GAP 2.0 uses a ''three-parent'' crossover operator and GAP 3.0 implements a ''population splitting'' scheme. Both of these new variants were introduced to enhance the ''mixing'' of schemata in the crossover operation, with the intent of generating many novel chromosome strings among the offspring. Each GA method was evaluated with respect to the sampling of conformational energy, ϕ and ψ torsional angle space for each residue, and convergence at each bit position for the ϕ and ψ torsional angles. As a prelude to optimizing GA performance for conformational search tasks, this analysis was undertaken to illustrate the sampling characteristics that arise from the use of different combinations of GA operators and parameters. In addition, the performance of these methods was compared to previous theoretical investigations of the conformational behaviour of [Met]-enkephalin.

Methods

GENERAL

Calculations were performed on IBM RS/6000 RISC workstations operating under AIX. Source code for the genetic algorithm programs was written using the ANSI FORTRAN 77 Standard. Minor

modifications to the ECEPP/2 source code [51] were made to allow compilation with the AIX XL Fortran compiler and to permit the passing of variables between ECEPP/2 and the GA subroutines. Data analysis was carried out in part with the commercial statistical analysis package SPLUS (MathSoft Inc., Seattle, WA). Energy minimization was performed using Powell's method.[52] To assess the GA methods' capabilities to sample the maximum available conformation space, electrostatic interactions were evaluated using a vacuum dielectric ($\epsilon = 1$) and an infinite cutoff. In the ECEPP/2 force field, bond angles and bond lengths are fixed at constant values and only torsional angles are permitted to vary.

The molecule used for this study was the pentapeptide [Met]-enkephalin (Tyr–Gly–Gly–Phe–Met) with both termini represented in their neutral forms; that is, —NH_2 and —COOH. This important biomolecule and its analogue [Leu]-enkephalin have been the subject of intense experimental and theoretical investigation. Therefore, this peptide provides a reliable standard upon which to base an assessment of a GA method's capability to explore conformation space.

IMPLEMENTATION OF GAP 2.0 AND GAP 3.0

The GAP 1.0, 2.0, and 3.0 programs were implemented and varied from each other in the initialization process as well as in the crossover operator. For input, each program required values for the population size, the mutation rate, the number of generations, a seed value required by a random number generator, and the necessary information for energy calculations using the ECEPP/2 force field (e.g., primary residue sequence). The initial parent population included 50 conformations in which all omega angles were set to 180° and all other angles were randomly generated. A total of four different starting populations were used in this study to gain an approximate assessment of the influence of the starting population on the outcome of each GA run. Each angle was represented by an eight-bit binary string thereby allowing angles that were multiples of $360/2^{8-1} \approx 1.41°$. Because there are 24 torsional angles in the ECEPP/2 description of [Met]-enkephalin, each conformer could be described by a 192-bit binary string chromosome. Schemata from parent conformations were recombined using one of three crossover operators: uniform crossover; ''three-parent'' crossover; and uniform crossover preceded by a ''population splitting'' scheme. In a previous study, uniform crossover was shown to be useful in preventing ''preconvergence.''[46] This operation was augmented in the three-parent crossover scheme by generating offspring from three as opposed to two parents. Initially, two-parent conformers were chosen to produce an intermediate offspring conformer. This intermediate was then recombined with a third parent to produce one final offspring conformer. This scheme was implemented to enhance the ability of the GA to generate conformers that were dissimilar to any of the parent conformers. Another approach to accomplish this was implemented in the ''population splitting'' scheme. This modification resulted in the division of the parent conformations into two groups of equal size. Parents from one group were permitted to recombine only with parents from the other group. This was done to decrease the likelihood that any single parent conformer could be involved in crossover with other parents of similar fitness and conformation. For each crossover operator, the parent conformer's fitness was not used in the crossover operation. This was done to maximize the variety of schemata—and consequently conformational diversity—that could appear among the offspring, given the small population size. In each GA program, the offspring were subjected to a mutation operator that randomly flipped bits at a specified mutation rate. For this study, a total of five different mutation rates were used: 0.00, 0.01, 0.03, 0.05, and 0.07. Once the number of offspring was equivalent to the number of parents—a total of 100 conformers—the chromosome binary string for each conformer was translated into real number torsional angle values that were then used to calculate the ECEPP/2 energy. The calculated energy was used as the fitness measure. To ensure that no two conformers were exactly the same, a ''diversity'' operator was implemented. An offspring was considered to be similar to a parent if more than half of its torsional angles were within 5° of the corresponding angles in a parent. If a similar parent–offspring pair was found, then the highest energy conformer of the pair was mutated at a high mutation rate. From the full population of 100 parents and offspring, the best performing (i.e., lowest energy) half of the population was selected for the next generation of parents. Each GA run was stopped after 1000 generations. Over the course of each run, the sampling of the ϕ and ψ torsional angle space for each residue was examined by binning the torsional angle values of each conformer every second generation. The average

population energy and the lowest conformer energy at each generation were also recorded, as was the usage of the diversity operator. The average value at each bit position was recorded at every second generation to observe convergence and schema propagation in the evolving parent conformer populations. Finally, each parent conformer in the final generation of each GA run was energy minimized and the conformers in the final populations were compared with low-energy conformations found in other studies.

Results

CONFORMATIONAL ANALYSIS OF [MET]-ENKEPHALIN CONFORMATIONAL SPACE

The [Met]-enkephalin pentapeptide is a highly flexible molecule. This was reflected by the multitude of structurally dissimilar low-energy conformers found in each GA run of this study. As expected, the peptide backbone displayed the most conformational variety at the glycine residues in the second and third position, with torsional angle values frequently appearing in each quadrant of the (ϕ, ψ) map covering $-180° < \phi < 180°$, $-180° < \psi < 180°$. The residues with large side chains showed a more restricted torsional angle range largely confined to the region spanned by $-180° < \phi < 0°, 0° < \psi < 180°$. For each GA method, regardless of the parameter settings and initial population used, the tyrosine, phenylalanine, and methionine residues each displayed the same unique conformational features in the peptide main chain.

At Tyrl, ϕ and ψ were found to exist mostly in the range $-180° < \phi < -30°$ and $120° < \psi < 180°$. This region encompasses many secondary structure motifs including parallel β-sheet, antiparallel β-sheet, and the collagen helix. Torsional angle values were also found to a much lesser extent for the right-handed α-helix and in the region $\phi \sim 70°, \psi \sim 180°$. (The latter conformational domain has not been commonly observed in the solution phase and its appearance here is likely to be permitted through the use of vacuum conditions.)

For both Gly2 and Gly3 there was great variance in ϕ, ψ-sampling as GA parameters were changed. This reflected the high degree of conformational flexibility afforded by these residues. For each glycine, torsional angle values were found in each quadrant, indicating the accessibility of both extended and coiled structures. These residues present numerous possibilities for both the β-turn and β'-turn motifs. In a previous study by Nayeem and coworkers,[53] the involvement of Gly3 in a type II′ β-turn structure has been reported as a key feature of the vacuum-phase global minimum-energy conformation. It is clear that the predominance of this structure is undermined by the inclusion of a glycine residue which permits accessibility to other conformational domains.

The conformational range of the peptide backbone was most restricted at Phe4. Torsional angles were found almost exclusively in the region $-180° < \phi < -80°, 120° < \psi < 180°$, indicating the preponderance of an extended conformation at this position. Although a very minor degree of sampling occurred in the right-handed α-helix and other regions of the (ϕ, ψ) map, the GA methods examined in this study fail to place this residue into a clear β-turn type structure.

At Met5, the greater conformational freedom available at the peptide terminus is reflected in the larger range for ψ than for ϕ. In most of the GA runs, both the right-handed α-helix and β-turn domains appear to be equally accessible to this residue. The β-turn region covered the area $-180° < \phi < -50°, 90° < \psi < 180°$, whereas the α-helix region spanned $-180° < \phi < -50°$ and $-80° < \psi < 0°$.

After energy minimization of the final parent conformer generations from each run, the torsional angle domains for tyrosine, phenylalanine, and methionine remained unaltered. For both glycine residues, torsional angle values also appeared in each quadrant of the (ϕ, ψ) map. The lowest energy conformer found in this study and the global minimum-energy conformation reported by Nayeem and coworkers[53] are shown in Figure 1a and b. Although there are many structural similarities between the two, the lowest energy conformer from this study was approximately 3 kcal/mol above the global energy minimum.

EVOLUTION OF CONFORMER ENERGY

For each GA method in this study, the average parent conformer energy and the lowest conformer energy were recorded at each generation (see Fig. 2). In each run, the average parent energy starts at a very high value and decreases rapidly within the first 100 generations. From 100 to 1000 generations, the average parent energy profile typically levels off to either a shallow downward slope or a plateau < 20 kcal/mol above the global minimum. For all methods, both the average parent energy and the

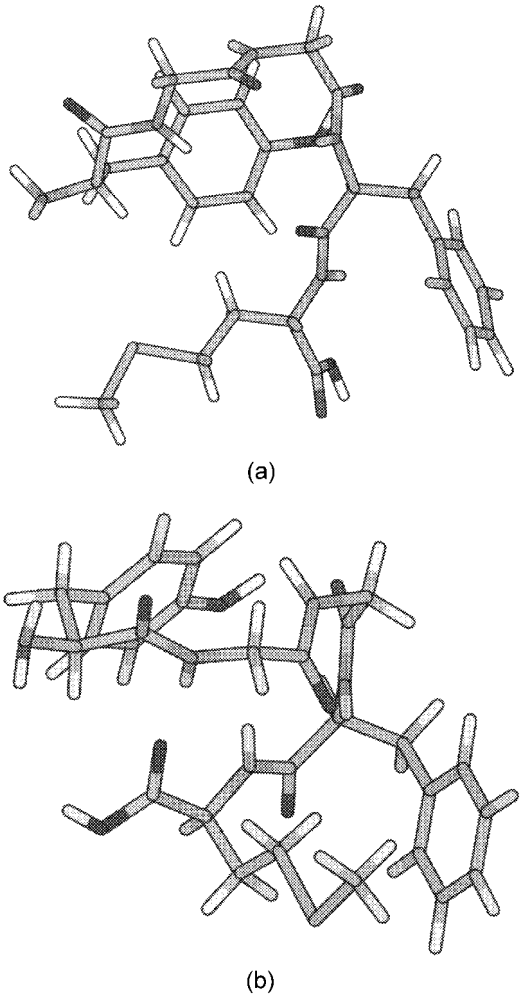

FIGURE 1. (a) Lowest energy conformer of [Met]-enkephalin found using GAP 1.0, 2.0, and 3.0. (b) Global minimum-energy conformer of [Met]-enkephalin reported by Nayeem et al.[53]

lowest conformer energy at the final generation decreased with decreasing mutation rate. This reflects the increased disruption of useful schema in the offspring by the increased occurrence of mutations. As the average parent energy decreases during a run, the usage of the diversity operator increases and eventually fluctuates around a constant level. Most of the mutants generated through the diversity operator were found among the offspring, reflecting the loss of diverse schema in the parent population at later generations. Moreover, the continued high use of the diversity operator at later stages of the run suggests that most mutations did not result in low-energy conformers that increased schema diversity in the subsequent parent population. The diversity operator was also associated with the mutation operator in that the greatest usage was observed at the lowest mutation rates (e.g., 0.0 or 0.01) and this usage decreased with increasing mutation rate. This reflects the ability of the mutation operator to enhance structural variety among the offspring conformers over the course of the run. These trends were common to all GA runs.

Whereas the average parent energy and the lowest conformer energy were strongly influenced by the mutation rate, the choice of crossover operator had little effect. For the three methods examined, large variances in the final generation energy statistics were observed over all mutation rates and initial populations (see Table I).

SAMPLING OF ϕ AND ψ TORSIONAL ANGLE SPACE

For runs less than 1000 generations, the ϕ and ψ angles of all residues in each parent and offspring conformer were recorded in a two-dimensional array of bins every second generation. Thus, for each residue, a total of 50,000 (ϕ, ψ) samples were taken from each 1000 generation run. Each bin corresponded to a $5° \times 5°$ area on the (ϕ, ψ) map yielding a total of $72^2 = 5184$ bins. If sampling was evenly distributed among all bins, one would expect to find ~ 10 samples per bin at the end of the run. For the GA methods in this study, sampling was characterized by the following quantities: the most frequent samples/bin value; the percentage of bins with 10 samples or less; the percentage of samples in all bins with 10 samples or less; and the percentage of samples drawn from each $90° \times 90°$ quadrant of the torsional angle map (where quadrant $1 \equiv -180° < \phi < 0°, 180° < \psi < 0°$; quadrant $2 \equiv -180° < \phi < 0°, 0° < \psi < 180°$; quadrant $3 \equiv 0° < \phi < 180°, 0° < \psi < 180$; and quadrant $4 \equiv 0° < \phi < 180°, -180° < \psi < 0°$). These measures permitted the assessment of the extent to which sampling was concentrated as well as the location of the most frequently sampled torsional angle regions.

As expected, sampling for both glycines was more widely distributed across the entire torsional angle map than for Tyrl, Phe4, and Met5. For each residue, sampling over the run was highly concentrated in small torsional angle ranges, as indicated by the large number of bins that contained only a few samples each. For this entire study, the most frequent samples/bin value ranged from zero to four. This indicated that the GA's capability to find new peptide backbone conformations became

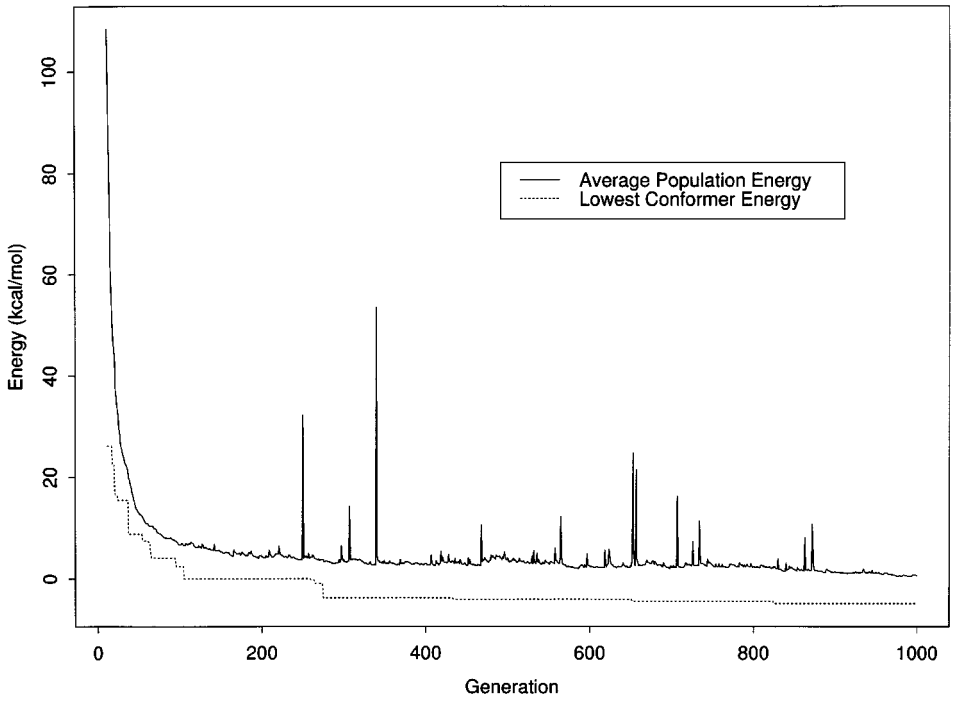

FIGURE 2. Lowest conformer energy and average conformer energy over 1000 generations from a representative GA run.

severely curtailed at an early point during the run. In all cases, regardless of the residue type, mutation rate, initial population, or crossover operator used, on average, 85–95% of all bins contained ten samples or less. However, the mutation rate had a surprising influence on other sampling characteristics for each residue. It was expected that, as mutation rate was increased, sampling should become more evenly distributed throughout the torsional angle map. However, this was not found to be the case. As the mutation rate was increased, the most frequent samples/bin value for all residues decreased from four to one or zero. Moreover, although the number of bins with ten or

TABLE I.
Lowest Conformer Energy (LCE) and Population Average Energy (PAE) for Each GAP Program at Mutation Rates 0.00, 0.01, 0.03, 0.05, and 0.07.[a]

	Mutation Rate				
	0.00	0.01	0.03	0.05	0.07
GAP 1.0 PAE	20.03 ± 11.14	3.32 ± 1.43	2.92 ± 0.52	4.14 ± 0.67	5.42 ± 0.78
GAP 1.0 LCE	−2.01 ± 1.52	−2.39 ± 0.76	−1.70 ± 0.93	0.92 ± 1.02	1.96 ± 0.57
GAP 2.0 PAE	5.03 ± 1.43	2.37 ± 1.36	2.47 ± 0.88	3.93 ± 0.83	4.57 ± 0.99
GAP 2.0 LCE	−2.51 ± 1.53	−2.57 ± 1.54	−0.89 ± 1.54	0.19 ± 1.38	1.81 ± 1.24
GAP 3.0 PAE	8.37 ± 4.92	3.09 ± 0.86	2.54 ± 1.05	3.88 ± 0.87	5.36 ± 1.29
GAP 3.0 LCE	−2.69 ± 1.18	−2.63 ± 1.27	−1.84 ± 1.41	0.33 ± 0.73	1.49 ± 2.25

[a] Values are averages from four runs; each run was started from a different initial population. All values are in kilocalories per mole.

fewer samples was similar for all runs, the percentage of samples found in these bins decreased as mutation rate increased. This meant that, with increasing mutation rate, more samples were found in fewer bins. For each residue, at mutation rate 0.00, approximately 55% of all samples were found in less than 13% of all bins. At mutation rate 0.07, 75–85% of all samples were found in the same number of bins. The mutation rate also affected sampling over the four quadrants. At low mutation rates, each quadrant contained at least 10% of all samples, whereas, at high mutation rates, some quadrants contained less than 3% of the total samples. This was noted in all residues, despite the unique sampling patterns seen in each (see Fig. 3). In Tyrl, sampling occurred predominantly in the second quadrant with some also occurring in the third quadrant. In both glycine residues, sampling in each quadrant was biased by the initial population and each quadrant was sampled to a different extent depending on the initial population used. In Phe4, sampling took place mostly in the second quadrant with the remaining three quadrants receiving the same number of samples. In Met5, quadrants one and two both accounted for most of the sampling. The extent to which each of these quadrants was sampled varied with the initial population. For all residues, as mutation rate increased, these patterns were exacerbated. For example, for uniform crossover alone with mutation rate 0.00, approximately 60% of the sampling for Gly2 took place in quadrant 1. At mutation rate 0.07, this increased to around 73%.

CONVERGENCE TRENDS AND SCHEMA PROPAGATION

The capability of a GA to create novel offspring conformations is dependent on the conformational diversity in the parent conformer population. This diversity can be assessed by counting the number of bit positions in the parent population that have converged to either one or zero. Because each bit position accounts for a different torsional angle range, the influence of each bit value will vary. For example, a bit which is found at the beginning of the eight bit string for a torsional angle will account for $2^{1-1} \cdot 1.41° = 1.41°$ of the torsional angle range, whereas a bit at the end of the string accounts for $2^{8-1} \cdot 1.41° \approx 180°$. The average value at a bit position reflects the proportion of the parent population that has converged. For example, an average value of 0.6 indicates that, out of 50 parent conformers, 30 have a ''1'' at the specified position and 20 have a ''0.'' When the average value approaches either 1.0 or 0.0 then the diversity at that bit position has essentially been lost and that position no longer plays a role in introducing conformational diversity through crossover. Alterations at that bit position in any conformer may then only occur through mutation. By tracking the average value at each bit position, it is possible to evaluate the loss of exploratory capability as well as the exploitation of fitness-improving schemata.

For each GAP version, the average bit value at each position in the peptide backbone was recorded. Regardless of the crossover operator

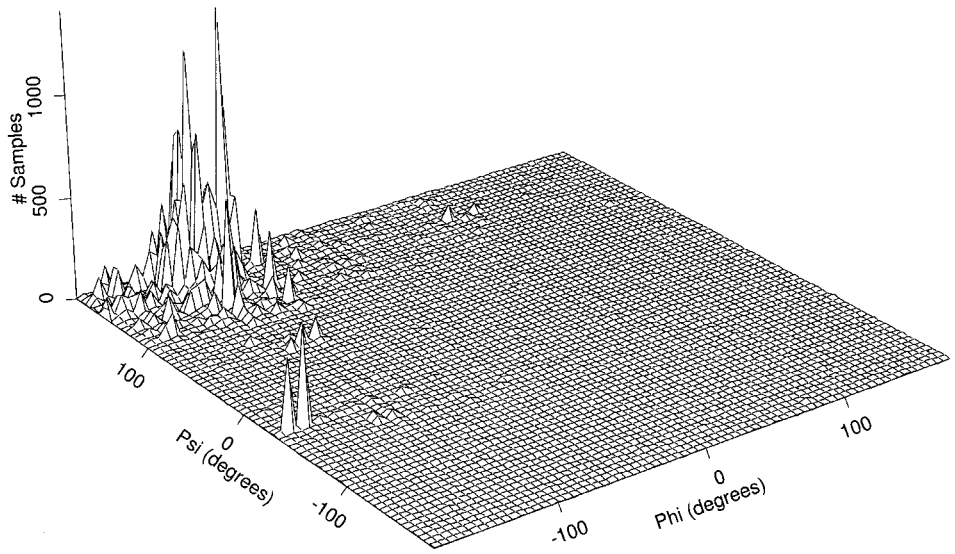

FIGURE 3. Sampling of Tyrl ϕ and ψ torsional angle space from a representative GA run.

used, the bit positions that accounted for large torsional angle ranges (i.e., 180°, 90°, 45°, and 22.5°) showed similar behavior for Tyrl, Phe4, and Met5 (see Fig. 4). For the Tyrl ϕ angle, the average value at the eighth bit (covering the 180° range) converged quickly to ≥ 0.98, indicating that only 1 out of 50 conformers had a zero bit value at this position. For Tyrl ψ, bit positions 7 and 8 converged quickly to > 0.8 and 0.0, respectively. This meant that, for this residue, the effective torsional angle range that was searched by the GA was restricted to $0° > \phi > -180°$ and $180° > \psi > 0°$. For both glycine residues, convergence in bit positions varied as either the initial population or the mutation rate was changed. In many cases, the same bit position converged to 1.0 in one run and 0.0 in another. Thus, although there was no consistent trend from run to run, many bit positions did converge nevertheless. Moreover, for all runs with mutation rate greater than 0.00, converged bits rarely showed any further change. Therefore, even for conformationally flexible residues like glycine, the GA quickly lost the capability to search the entire torsional angle range in either ϕ or ψ. For Phe4 ϕ, convergence to 1.0 in position 8 occurred very quickly in all runs. The seventh bit showed a trend toward a low average value of < 0.2 and, in most cases, converged to 0.0 before the halfway mark of the run. This resulted in restricted ϕ sampling mostly in the range $-180° < \phi < -90°$. The ψ angle displayed a strong tendency for convergence to 0.0 in the eighth bit, whereas the positions from 5 to 7 also displayed consistent trends, which became apparent in the early stages of the run. Bit position 5 showed a trend toward an average value of < 0.5, whereas the sixth and seventh positions converged in most cases to 1.0. Consequently, sampling in ψ was mainly restricted to $135° < \psi < 180°$. For all runs, this residue possessed the highest number of converged bit positions. For Met5, the range in ϕ became restricted to values between $-180°$ and 0° very quickly, but in ψ only the seventh bit position showed a propensity to converge toward 1.0, whereas the remaining bit positions possessed average values in the range 0.2–0.8. This permitted the GA to search two conformation space regions with the same ϕ range, but different ψ ranges: $90° < \psi < 180$ and $-90° < \psi < 0°$.

Convergence in a chromosome population indicates the propagation of good schema as well as the possible entrapment of the GA in a local minimum. For a finite-sized population, the latter is a realistic concern and is likely to diminish the efficiency of a GA search. The number of converged bit positions in the final conformer populations

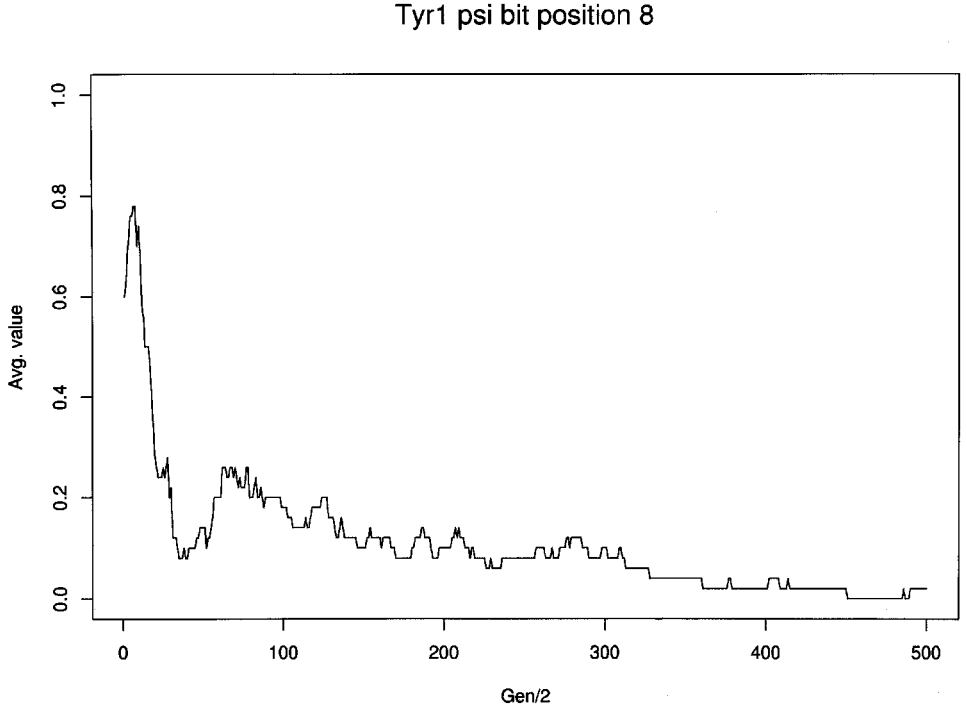

FIGURE 4. Average value at bit position 8 for Tyrl ψ from a representative GA run.

TABLE II.
Number of Converged Bits at Each Mutation Rate for Each GAP Program.[a]

GA Program	Mutation Rate 0.00	0.01	0.03	0.05	0.07
GAP 1.0	100 ± 16	97 ± 7	87 ± 2	82 ± 3	77 ± 5
GAP2.0	121 ± 5	111 ± 2	91 ± 1	81 ± 3	83 ± 5
GAP3.0	105 ± 7	101 ± 3	88 ± 2	81 ± 2	78 ± 4

[a] Values are averages from four runs; each run was started from a different initial population.

was only slightly affected by the use of different crossover operators (see Table II). At each mutation rate, both GAP 1.0 and 3.0 showed slightly fewer converged bits than GAP 2.0. For each GAP version, the number of converged bits declined with increasing mutation rate. Within the chromosome binary string, 40 bit positions are used to describe fixed omega angles leaving 152 bits that can vary through crossover. For GAP 1.0, the percentage of variable bit positions that converged to either > 0.9 or < 0.1 varied from about 39% at mutation rate 0.00 to 31% at mutation rate 0.07. This was also the case with GAP 3.0. GAP 2.0 showed convergence in approximately 53% of the bits at mutation rate 0.00 and about 33% at mutation rate 0.07. For each program, convergence occurred mainly in the last four bit positions of the side-chain torsional angles of Tyrl, Phe4, and Met5. These bit positions corresponded to increments of 22.5°, 45°, 90°, and 180°. This meant that the torsional angle range that was searched for these residues' side chains was $< 180°$ and in some cases $< 90°$. This behavior suggests that each GAP program is susceptible to entrapment in local minima at all mutation rates.

For each GAP version, many bit positions' average values were highly correlated to each other over the run (see Table III). This suggested either the simultaneous propagation of many short schemata or the propagation of fewer long schemata. In GAP 1.0, virtually no correlations with $r > 0.90$ were observed at mutation rate 0.00. Upon examining the profiles for average bit values, it appears that the high usage of the diversity operator lead to large random fluctuations over the run. With the introduction of a mutation rate of > 0.00, this disruption was removed. For each of the mutation rates 0.01 to 0.07, approximately 20–25 positions show correlation to at least one other position with $r > 0.90$. The separation between these correlated positions was seen to vary from one (i.e., adjacent bit positions) to over 160 (i.e., over the entire span of the chromosome). Groups of correlated bits varied in size from 2 to over > 10. Correlation appeared in every residue and in virtually every torsional angle suggesting the existence of useful schemata that could span many residues. For lower mutation rates, the absence of correlations does not result in fewer converged bits or in a higher average population energy. This suggests that, in [Met]-enkephalin, higher order schemata are built from lower order ones. The bit positions that were most often involved were located in the last four positions for each angle. In contrast to GAP 1.0, many correlated bits appeared at mutation rate 0.00 for GAP 2.0, and this number (between 40 and 50 bit positions) changed little as the mutation rate was increased. For GAP 3.0, about 20 correlations appear at mutation rate 0.00 and 40–50 appear at muta-

TABLE III.
Number of Correlated Bit Positions at Each Mutation Rate for Each GAP Program.[a]

GA Program	Mutation Rate 0.00	0.01	0.03	0.05	0.07
GAP 1.0	1 ± 2	24 ± 4	20 ± 5	21 ± 8	24 ± 5
GAP 2.0	47 ± 6	48 ± 6	51 ± 7	53 ± 4	54 ± 3
GAP 3.0	22 ± 1	41 ± 5	51 ± 9	53 ± 7	43 ± 8

[a] Values are averaged from four runs; each run was started from a different initial population.

tion rates 0.01−0.07. For both of these programs the diversity operator was used less frequently than in GAP 1.0, especially at low mutation rates (0.00 and 0.01). This suggests that the diversity operator has a disrupting effect on the propagation of high-order schemata. At the same time, this disruption does not appear to have a negative effect on the evolution of population energies.

Discussion

EVALUATION OF [MET]-ENKEPHALIN CONFORMATION SPACE BY *GA*

The pentapeptide [Met]-enkephalin and its analogue [Leu]-enkephalin present a prototypical case of the difficulty in elucidating biologically relevant conformations. The physiological effects of these highly potent compounds are mediated by multiple opioid receptors.[54] These receptor subtypes play a shared role in many biological functions, but they are also involved in separate processes in both normal and pathological states. The elucidation of the pharmcophoric and toxicophoric elements of the enkephalin peptides has been confounded by their tremendous conformational diversity, as noted in many previous experimental structural analyses. Although three crystal structures of [Met]-enkephalin have been elucidated, each shows the influence of intermolecular hydrogen bonding resulting in an extended conformation.[3−5] In contrast, two-dimensional NMR studies reveal a propensity for a coiled structure, although there is no clear indication of a predominant solution-phase conformation.[6,7] In this study, three GA methods were assessed for their ability to search the [Met]-enkephalin peptide conformation space in the absence of *a priori* information. Each method was capable of finding many structurally diverse low-energy conformers in a relatively short time. Among these conformers, the peptide backbone features found in the global minimum energy structure were frequently found. Upon energy minimization of the final generation of conformers from each GA run, many structures were found to be within < 10 kcal/mol of the global minimum energy. For this series of vacuum phase calculations, a variety of both extended and coiled conformations were found at similar energy values. It was also clear from the flexibility of Gly2 and Gly3 that an accurate portrayal of the conformational behavior of this molecule must be constructed from a large ensemble of low-energy structures. In contrast, the flexibility in the peptide backbone was greatly restricted at Tyrl, Phe4, and Met5, which led to the observed conformational preferences of the residues. The GA also suggested the dependence of residue flexibility on the position in the primary sequence. Therefore, although tyrosine and phenylalanine have similar side chains, the terminal position of Tyrl affords greater backbone flexibility than in Phe4. The torsional angle ranges of these residues encompass most previously found values for low-energy [Met]-enkephalin conformers. It is also of interest to note that Met5 displayed two conformational domains at low energy and that these domains were established at early stages in each GA run. This suggests that, at low energy, this residue has an equal probability of assuming one of two structural motifs. This presents a variable for peptidomimetic design in which one portion of an analogue series is constructed to display either a helical or extended conformation at this position. As expected, comparison of different GA runs showed that Gly2 and Gly3 afforded a diverse conformer population at low energy. In the absence of *a priori* information, the structure at these positions also presents another design variable. For [Met]-enkephalin, it appears that a G−G type II′ β-turn is conducive to binding at the μ-opioid receptor,[55] although this does not preclude the importance of other conformations.

CROSSOVER IN *GA*

The use of different crossover operators in this study had little effect on the sampling characteristics examined. For the same set of initial parameters (mutation rate, population size, initial population), each GAP program required approximately the same amount of time to complete a 1000-generation run. Each program was also similarly affected by changes in the mutation rate and initial population. Although each crossover operator influenced schema propagation, this did not appear to yield significant differences in the sampling of torsional angle space or in the evolution of either the average or lowest conformer energy. This suggests that the manner in which schemata are recombined from the parent conformers is not a vital contributor to the search mechanism of any of these programs. This is consistent with the conclusions drawn by van Kampen and coworkers[56] who noted that the recombination operator is not always a useful component of stochastic optimization strategies. Although crossover operations do transfer useful schemata from parents to offspring

in the early stages of the GA search, for highly fit populations this operator is redundant. For the conformational search of peptides, crossover operations tend to produce offspring conformers that are: (a) very similar to the parent conformers (more than half of the torsional angles are within 5° of the parent value); and (b) at higher energy than the parents. When offspring are similar to parents, the subsequent mutation from the diversity operator is so severe that the mutant is almost always at higher energy than any of the current parent conformers. In these cases, the crossover operation is wasted. Because any recombination of material is inefficient for a population of low-energy conformers, alterations in the crossover mechanism are unlikely to have a great effect on the progress of the GA toward optimal solutions. Therefore, in a population of low-energy conformers the role of the diversity and selection operators becomes far more important for both exploring new conformation space and for exploiting the previous sampling history.

COMPARISON TO OTHER CONFORMATIONAL SEARCH METHODS

The computational chemistry literature abounds with reports of conformational analyses for [Met]- and [Leu]-enkephalin. Of these, the comparative study of the simulated-annealing (SA) and Monte-Carlo-with-minimization (MCM) approaches by Nayeem and coworkers provides a well-defined [Met]-enkephalin structure at the apparent global energy minimum in the absence of water.[53] From that study it was reported that a type II′ β-turn structure involving Gly3 and Phe4 was observed at the global minimum of -12.9 kcal/mol. Ishida and coworkers[57] applied an MD approach to the study of folding within the [Met]-enkephalin peptide and reported a type II′ β-turn structure over Gly2 and Gly3 as being the most likely equilibrium conformation. It is of interest to note that this motif has also been proposed as a conformational determinant for binding at the μ-opioid receptor.[58] Meirovitch and Meirovitch[59–61] have applied both the MCM method and a modification—the free-energy Monte-Carlo-with-minimization procedure (FMCM)—to the elucidation of ''local microstates'' above the global potential energy minimum and the global harmonic free-energy minimum. It was shown that the energy range within 7.5 kcal/mol of the global energy minimum could easily accommodate thousands of diverse conformers differing by 50° in at least one torsional angle.

The GA codes examined do not perform as well as other methods for finding low-energy conformer populations. The FMCM approach, the MCM approach, and SA all show better efficiency at finding the global energy minimum structure and in finding low-energy conformation space. Each GA code suffers from the same drawback: an inability to find novel useful schema after the population is at low energy. Each is also susceptible to the same variance from changes in the initial population and mutation rate. Although convergence at many bit positions was affected upon altering the crossover operator, these differences did not translate to large differences in torsional angle sampling characteristics or the evolution of conformer energies. The ineffectiveness of the GAP programs in improving low-energy conformer populations can be associated with poor exploratory capability. For example, bit positions corresponding to the ϕ and ψ angles for Gly2 and Gly3 were found to have converged in many runs, although this is not warranted. Upon comparison of GA runs which differ only in the initial population used, it is clear that these convergence trends illustrate the inability of each algorithm to remove the bias introduced by the starting population. This is due to the finite size of the conformer population and the limit of 1000 generations on each GA run.

FURTHER DEVELOPMENTS

Alterations to the selection operator should be investigated. In the current GAP programs, the crossover operation is wasted during the later stages of the run because most of the offspring that are generated are at a higher energy than any of the existing parent conformers; these offspring are discarded and the crossover operation has no net effect. Consequently, novel useful schema, which may be ''masked'' in an offspring conformer, will not be given a chance to be incorporated into low-energy conformers at later generations. One approach toward dealing with this is to permit the survival of some high-energy conformers by implementing a probabilistic selection criterion. For example, if an offspring conformer is at a lower energy than the previous average population energy, then it is always accepted. If it is at a higher energy than the previous average then it has an acceptance probability that is determined in some way; for instance, through application of a Metropolis criterion. This should permit explo-

ration of torsional angle space that is removed from low-energy regions.

As an alternative to the binary representation used in the GAP programs, real number encoding permits more precise control over torsional angle values. Additionally, the incorporation of torsional angle constraints may also be simpler using real rather than binary physical variables. However, the reduced number of schemata available in a real number representation may diminish the GA's ability to find novel solutions.

In the present study, the choice of [Met]-enkephalin as the sole test case illustrated many of the differences (and similarities) between the three GAP variants. The subsequent optimization of GA parameters and operators will require the analysis of a large and diverse set of molecules. Because GAP 1.0, 2.0, and 3.0 rely on the ECEPP/2 force field for energy calculations, the use of these programs can be generalized to a wide array of linear peptides. In principal, the GAP programs may also be modified for use with other force fields to permit the examination of nonpeptide molecules.

Conclusions

Each genetic algorithm program in this study was proficient at quickly finding low-energy conformation space for [Met]-enkephalin in the absence of *a priori* structural knowledge. Within approximately 10 kcal/mol of the reported global minimum energy structure, a wide variety of dissimilar conformers was found. Among these were many structures which represented secondary structure motifs including: the right-handed α-helix; several types of β-turn, including the type II′ β-turn; and the extended β-sheet conformation. Each GAP program indicated the broad flexibility at Gly2 and Gly3, in contrast to the conformationally restricted peptide backbone at Tyrl, Phe4, and Met5. Furthermore, ϕ and ψ angles for Met5 appeared in two groupings corresponding to the β-sheet and right-handed α-helix regions. Conformational features found in many previously reported low-energy conformations—including the putative μ-opioid receptor-bound conformation—appeared frequently throughout the GA runs.

Through this preliminary study the performance levels of GAP 1.0, 2.0, and 3.0 were compared with respect to both energy- and structure-based criteria. At low mutation rates (0.00 and 0.01), both GAP 2.0 and 3.0 displayed the best energy minimization capability, as was evident from the lowest conformer energies in Table I. Finding low-energy conformer populations was best served by GAP 2.0 at mutation rate 0.01; from four populations, the mean population average energy was 2.37 kcal/mol. Although sampling patterns were equivalent for all programs, structural properties based on schema propagation provide an additional basis for comparison. The presence of converged bits permits identification of torsional angle ranges that are conducive to low-energy conformations. At each mutation rate, GAP 2.0 generated as many or more converged bit positions than either GAP 1.0 or 3.0; the highest number of converged bit positions was found using a mutation rate of 0.00. Thus, GAP 2.0 was best at divulging conformational restrictions that were compatible with low energy. Although correlation among bit positions is not necessary to ensure the survival of high-order schema, the time-dependent propagation of schema provides an indication of their viability. Thus, for example, high-order schema seen to propagate quickly indicate their feasibility irrespective of the presence of other schema. The greatest number of correlated bit positions was seen at mutation rate 0.07 using GAP 2.0, from which an average of 54 ± 3 correlated bits were observed. Thus, GAP 2.0 was best at revealing the possible interdependence between different torsional angles.

Although the elucidation of low-energy conformer space was relatively straightforward, subsequent improvements in conformer populations occurred very slowly. The operators implemented in this study were not effective for searching conformer space after a low-energy region had been found. The diversity operator prevented convergence but its high usage did not aid efficient exploration. Although virtually all of the late generation offspring were severely mutated through the diversity operator, improvements in both the lowest conformer energy and the average population energy were infrequent. Each of the crossover operators permit high-order schema to propagate with varying degrees of disruption. The GAP 1.0 program showed greater use of the diversity operator than GAP 2.0 and GAP 3.0, resulting in more changes at converged bit positions. However, the GAP programs showed only minor differences in torsional angle sampling patterns. All GAP programs were susceptible to changes in the initial population. Also, mutation rate had a similar effect on all programs in that low mutation rates were associated with: frequent occurrence of struc-

turally similar parent-offspring pairs; low average population energy; low best conformer energy; and more evenly distributed sampling of ϕ and ψ angles in all residues.

The initial outlook for genetic algorithms in conformational search tasks appears promising but there is clearly much room for improvement. The incorporation of more sophisticated sampling strategies, such as niching, probabilistic selection, and elitism present possible routes for dealing with the obstacle of finite population size. The possible hybridization of the GA with energy minimization routines should also be explored, although this could incur a large increase in the required computation time. This strategy permits the exploitation of conformational data in populations that have not improved after many generations.

Acknowledgments

The authors thank Oreola Donini, Dr. Heather L. Gordon, and Mark N. Anderson for many useful discussions throughout this study.

References

1. Piela, L.; Scheraga, H. A. Biopolymers 1987, 26, s33–s58.
2. Scheraga, H. S. In: Lipkowitz, K. B.; Boyd, D. B., eds. Reviews in Computational Chemistry, Vol. 3; VCH: New York, 1992; p 73.
3. Mastropaola, D.; Camerman, A.; Camerman, N. Biochem Biophys Res Commun 1986, 134, 698–703.
4. Doi, M.; Tanaka, M.; Ishida, T.; Inoue, M.; Fujiwara, T.; Tomita, K.; Kimura, T.; Sakakibara, S.; Sheldrick, G. M. J Biochem 1987, 101, 485–490.
5. Griffin, J. F.; Langs, D. A.; Smith, G. D.; Blundell, T. L.; Tickle, I. J.; Bedarkar, S. Proc Nat Acad Sci USA 1986, 83, 3272–3276.
6. Motta, A.; Tancredi, T.; Temussi, P. A. FEBS Lett 1987, 215, 215–218.
7. Graham, W. H.; Carter II, E. S.; Hicks, R. P. Biopolymers 1992, 32, 1755–1764.
8. Holland, J. H. Adaptation in Natural and Artificial Systems; MIT Press: Cambridge, MA, 1992.
9. Goldberg, D. E. Genetic Algorithms in Search, Optimization, and Machine Learning; Addison-Wesley, Reading, MA, 1989.
10. Beasley, D.; Bull, D. R.; Martin, R. R. University Comput 1993, 15, 58–69.
11. Beasley, D.; Bull, D. R.; Martin, R. R. University Comput 1993, 15, 170–181.
12. Jefferson, M. F. Pendleton, N.; Lucas, S. B.; Horan, M. A. Cancer, 1997, 79, 1338–1442.
13. Li, L.; Darden, T. A.; Freedman, S. J.; Furie, B. C.; Furie, B.; Baleja, J. D.; Smith, H.; Hiskey, R. G.; Pedersen, L. G. Biochemistry 1997, 36, 2132–2138.
14. Judson, R. In: Lipkowitz, K. B.: Boyd, D. B. eds. Reviews in Computational Chemistry, Vol. 10; VCH: New York, 1997, p 1.
15. Lucasius, C. B.; Kateman, G. Chemometrics Intell Lab Syst 1993, 19, 1–33.
16. Hibbert, D. B. Chemometrics Intell Lab Syst 1993, 19, 277–293.
17. Clark, D. E.; Westhead, D. R. J Comput Aid Molec Des 1996, 10, 337–358.
18. Maddox, J. Nature 1995, 376, 209.
19. Lucasius, C. B.; Kateman, G. Trends Anal Chem 1991, 10, 254.
20. Bangalore, A. S.; Shaffer, R. E.; Small, G. W. Anal Chem 1996, 68, 4200–4212.
21. Yokobayashi, Y.; Ikebukuro, K.; McNiven, S.; Karube, I. J. Chem Soc Perkin Trans 1996, 1, 2435–2437.
22. Rabow, A. A.; Scheraga, H. A. Prot Sci 1996, 5, 1800–1815.
23. Sun, S. Biophys J 1995, 69, 340–355.
24. May, A. C. W.; Johnson, M. S. Prot Eng 1995, 8, 873–882.
25. Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. A. J Molec Biol 1997, 265, 445–464.
26. Dandekar, T.; Argos, P. J Molec Biol 1996, 256, 645–550.
27. Gunn, J. R. J Chem Phys 1997, 106, 4270–4281.
28. Pedersen, J. T.; Moult, J. Curr Opin Struct Biol 1996, 6, 227–231.
29. Unger, R.; Moult, J. J Molec Biol 1993, 23, 75–81.
30. Dandekar, T.; Argos, P. Prot Eng 1992, 5, 637–645.
31. Dandekar, T.; Argos, P. J Molec Biol 1994, 236, 844–861.
32. Tuffery, P.; Etchebest, C.; Hazout, S.; Lavery, R. J Comput Chem 1993, 14, 790–798.
33. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. J Molec Biol 1997, 267, 727–748.
34. Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M. J Mol Struct (Theochem) 1994, 308, 191.
35. Walters, D. E.; Hinds, R. M. J Med Chem 1994, 37, 2527.
36. Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. J Comput-Aid Molec Des 1995, 9, 113.
37. Ring, C. S.; Cohen, F. E. Israel J Chem 1994, 34, 245–252.
38. Mestres, J.; Scuseria, G. E. J Comput Chem 1995, 16, 729–742.
39. Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. J Comput Chem 1993, 14, 1407–1414.
40. Judson, R. S.; Colvin, M. E.; Meza, J. C.; Huffer, A.; Gutierrez, D. Int Quantum Chem 1992, 44, 277–290.
41. McGarrah, D. B.; Judson, R. S. J Comput Chem 1993, 14, 1385–1395.
42. Brodmeier, T.; Pretsch, E. J Comput Chem 1994, 15, 588–595.
43. Niesse, J. A.; Mayne, H. R. Chem Phys Lett 1996, 261, 576–582.
44. Meza, J. C.; Judson, R. S.; Faulkner, T. R.; Treasurywala, A. M. J Comput Chem 1996, 17, 1142–1451.
45. van Batenburg, F. H. D.; Gultyaev, A. P.; Pleu, C. W. A. J Theor Biol 1995, 174, 269–280.
46. Jin, A. Y.; Leung, F. Y.; Weaver, D. F. J Comput Chem 1997, 18, 1971–1984.

47. Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. J Phys Chem 1975, 79, 2361–2381.
48. Némethy, G.; Pottle, M. S.; Scheraga, H. A. J Phys Chem 1983, 87, 1883–1887.
49. Sippl, M. J.; Némethy, G.; Scheraga, H. A. J Phys Chem 1984, 88, 6231–6233.
50. Némethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. J Phys Chem 1992, 96, 6472–6484.
51. ECEPP/2: Empirical Conformation Energy Program for Peptides (QCPE Program No. 454); Cornell University: Ithaca, NY.
52. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. Numerical Recipes in FORTRAN, 2nd Ed.; Cambridge University Press: Cambridge, 1992, p 387.
53. Nayeem, A.; Vila, J.; Scheraga, H. A. J Comput Chem 1991, 12, 594–605.
54. Simon, E. J.; Hiller, J. M.; Siegel, G. J.; Agranoff, B. W.; Albers, R. W.; Molinoff, P. B. Basic Neurochemistry, 5th Ed.; Raven: New York, 1994, p 321.
55. Loew, G. H.; Burt, S. K. Proc Nat Acad Sci USA 1978, 75, 7–11.
56. van Kampen, A. H. C.; Buydens, L. M. C. Chemometrics Intell Lab Syst 1997, 36, 141–152.
57. Ishida, T.; Yoneda, S.; Doi, M.; Inoue, M.; Kitamura, K. Biochem J 1988, 255, 621–628.
58. Loew, G. H.; Hashimoto, G.; Williamson, L.; Burt, S.; Anderson, W. Molec Pharmacol 1982, 22, 2667.
59. Meirovitch, H.; Vàsquez, J Molec Struct (Theochem) 1997, 398, 517–522.
60. Meirovitch, H.; Meirovitch, E. J Comput Chem 1997, 18, 240–253.
61. Meirovitch, E.; Meirovitch, H. Biopolymers 1996, 38, 69–88.